Robustness of HMM-based Speech Synthesis

Junichi Yamagishi[1], *Zhenhua Ling*[1,2], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom
[2]iFlytek Speech Lab., University of Science and Technology of China, Anhui, China
jyamagis@inf.ed.ac.uk, zhling@ustc.edu, Simon.King@ed.ac.uk

Abstract

As speech synthesis techniques become more advanced, we are able to consider building high-quality voices from data collected outside the usual highly-controlled recording studio environment. This presents new challenges that are not present in conventional text-to-speech synthesis: the available speech data are not perfectly clean, the recording conditions are not consistent, and/or the phonetic balance of the material is not ideal. Although a clear picture of the performance of various speech synthesis techniques (e.g., concatenative, HMM-based or hybrid) under good conditions is provided by the Blizzard Challenge, it is not well understood how robust these algorithms are to less favourable conditions. In this paper, we analyse the performance of several speech synthesis methods under such conditions. This is, as far as we know, a new research topic: "Robust speech synthesis." As a consequence of our investigations, we propose a new robust training method for the HMM-based speech synthesis in for use with speech data collected in unfavourable conditions.

Index Terms: speech synthesis, HMM, unit selection, HTS

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] has become established and well-studied, and is able to generate natural-sounding synthetic speech. In this method, the spectrum, fundamental frequency and segment duration are modelled and generated simultaneously within a unified HMM framework. A significant advantage of this model-based parametric approach is that speech synthesis is far more flexible compared with conventional unit-selection / waveform concatenation methods, since many model adaptation and model interpolation methods can be used to control the model parameters and thus the characteristics of the generated speech [2, 3]. In the current work, we demonstrate yet another advantage of the HMM approach: "robustness."

The ability to create synthetic speech with diverse speaker characteristics has many potential attractive commercial applications, such as virtual celebrity actors [4], as well as clinical applications such as synthetic replacement voices. The ability to create speech with the characteristics of a particular speaker could be combined with spoken language translation, to personalise speech-to-speech translation: a user's speech in one language can be used to produce corresponding speech in another language, while continuing to sound like the user's voice[1]. This technology would also have applications in dubbing foreign-language television programmes or movies.

In many of these applications, the available speech for the target speaker will suffer from noise or fluctuations in the recording conditions (changes in environment, microphone type and placement, etc.); this would be expected to significantly degrade the quality of the synthetic speech. Moreover, such "found" speech is unlikely to be phonetically balanced and will therefore lack some essential acoustic units. This causes severe problems in some systems: for example, concatenative systems must back off to some other unit, which may or may not sound acceptable. To realise these applications of speech synthesis, we start with an analysis of the performance of current speech synthesis methods on such imperfect data.

[1]See the *EMIME* Project. http://www.emime.org/

2. Systems

In our experiments, we compared the three major competing Text-to-Speech (TTS) methods. We will use the term *target speaker* to mean the speaker whose speech the synthesiser must reproduce; there may also be one or more *source speakers* whose speech may be used to build some of the systems under investigation.

The first method is the dominant, established, well-studied technique: unit-selection. This method concatenates units of speech, selected from a corpus of the target speaker's speech, to create new utterances [5]. The second is often termed "statistical parametric synthesis," in which a statistical model (usually a HMM) is trained on, or adapted to, the target speaker's speech. The third method is a hybrid of the statistical parametric and unit-selection techniques [6, 7], which has been shown to generate very natural-sounding synthetic speech when clean speech data are available for the target speaker [8].

2.1. Unit-Selection System

Festival [9] is a popular unit-selection speech synthesis system. We used Festival's "Multisyn" module [10], which provides a flexible, general implementation of unit selection and a set of associated voice building tools.

2.2. Statistical Parametric Systems

The HTS-2007 system [11, 12] is a high-quality speaker-adaptive HMM-based speech synthesis system developed by Nagoya Institute Technology and CSTR. In this system (Fig. 1), an average voice model using context-dependent multi-stream MSD-HSMMs is created using speaker-adaptive training on more than 10 hours of speech data uttered by many source speakers. This model is then adapted using speech data from the target speaker using a combined algorithm of constrained structural maximum a posteriori linear regression (CSMAPLR) [13] and maximum a posteriori (MAP) adaptation. The acoustic features for the MSD-HSMMs are three kinds of parameters required for a high-quality speech vocoding method with mixed-band excitation called *STRAIGHT* [14]: mel-cepstrum, $\log F_0$,

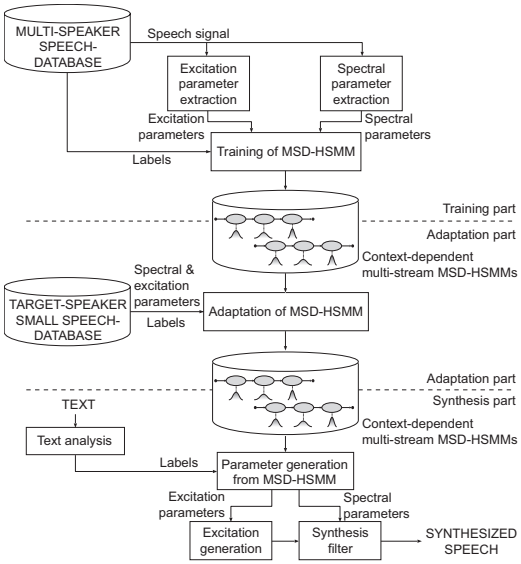


Figure 1: Overview of the HTS-2007 speech synthesis system.

and aperiodicity measures. Speech parameters are generated from the adapted MSD-HSMMs using a penalised maximum likelihood method [15].

Since the average voice models can utilise a large-scale data-rich speech database, and because both spectral and prosodic features such as $\log F_0$ and duration can be statistically and simultaneously transformed from the average voice model into those of the target speaker, we can robustly create voices even when only a relatively small amount of speech data is available for the target speaker. The synthetic speech generated from this system has a somewhat vocoded quality, although less so than earlier HMM-based speech synthesisers. Parts of this system have already been released in an open-source software toolkit called HTS ("H Triple S": HMM-based Speech Synthesis System) [16].

The HTS-USTC speech synthesis system [8] is also HMM-based, with context-dependent HMMs for spectrum, $\log F_0$ and phone duration being estimated from a single speaker database. There are three principal differences between HTS-USTC and HTS-2007: 1) HTS-USTC uses a MGE criterion [17] whereas HTS-2007 uses the ML criterion; 2) HTS-USTC uses line spectral pair (LSP) features whereas HTS-2007 uses mel-cepstrum features to represent the spectrum; 3) HTS-USTC only uses data from the target speaker, whereas HTS-2007 is speaker-adaptive.

2.3. Hybrid System

The HTS-USTC and iFlytek systems [8] use the same underlying HMMs but different waveform generation methods. In the HTS-USTC system, speech parameters are generated directly from the statistical models using a parametric synthesiser to reconstruct the speech waveform. On the other hand, the iFlytek system adopts a waveform concatenation method, in which a maximum likelihood criterion of the statistical models guides the selection of phone-sized candidate units from a single-speaker database [6, 7]. Both systems only use data from the target speaker.

3. Experiment

3.1. Data

We wished to separate the effects of two aspects of less-than-ideal data for the target speaker: lack of quantity / phonetic balance vs. suboptimal recording conditions, so we built voices for each of our four synthesisers using sets of sentences taken from two corpora. In all experiments, only target speaker data from the chosen subset was used to build the voice. For example, we did not utilise the full data set to train acoustic models used for segmentation, when building voices on the smaller sets. Note that the speaker-adaptive HTS-2007 system was trained on a substantial amount of clean speech data from other speakers, then adapted using the chosen subset of data from the target speaker.

The first corpus contains high-quality, clean speech data collected under controlled recording studio conditions by a male British English speaker with a received pronunciation (RP) accent. Subsets consisting of 768 randomly chosen sentences (about one hour in duration), 3063 randomly chosen sentences (about 4 hours in duration) and 6691 randomly chosen sentences (about 9.5 hours in duration) were used.

The second corpus consists of noisy data for comparison and was constructed from audio, freely available on the web, of a well-known American politician. These data were not recorded in a studio and have a small amount of background noise. The recording condition of the data is not consistent: the environment and microphone may vary. Subsets consisting of 978 randomly chosen sentences (about one hour in duration) and 3846 randomly chosen sentences (about four hours in duration) were used. For details of this data, please see [4].

3.1.1. Parameterisation

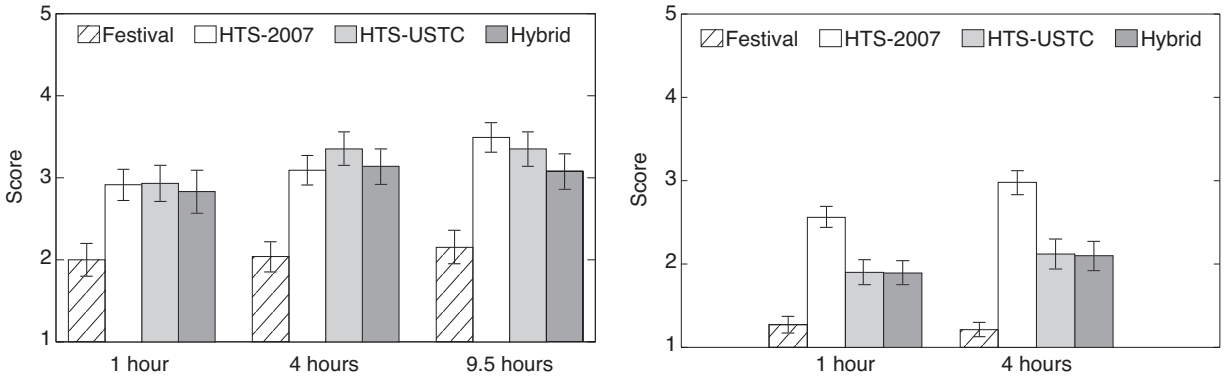
Speech signals were sampled at 16 kHz. F_0 for use in all synthesis methods was estimated using a voting method which combines the IFAS algorithm [18], a fixed-point analysis called TEMPO [19], and the ESPS get_f0 [20] tool. Voting reduces errors such as F_0 halving and doubling, and voiced/unvoiced errors. The spectral analysis methods varied according to system: Festival uses 13 MFCC coefficients (in the join cost), HTS-2007 uses 40 mel-cepstral coefficients, HTS-USTC uses 40 LSP coefficients, and the iFlytek hybrid system uses 13 mel-cepstral coefficients.

3.1.2. Labels and Lexicon

In order to exclude differences in front-end text-processing, we used the same labels and lexicon for the voice building and test sentence synthesis in all systems. The labels were generated using Unilex [21] and Festival's Multisyn module. Likewise, the same question set for the clustering of context-dependent HMMs was used in the HTS-2007, HTS-USTC, and iFlytek hybrid systems.

3.2. Results

All the systems were used to synthesise the story "The Little Girl and the Wolf" by James Thurber and the fairy tale "Goldilocks and the Three Bears." Neither of these texts were in the training data. The stories were split up into 12 and 22 utterances, respectively. In the "Little Girl" story, each utterance consisted of a single sentence, whereas each utterance consisted of two sentences in the "Goldilocks" story. 55 subjects (of which 47 were native speakers) were presented with syn-



(a) Clean data (b) Noisy data

Figure 2: Subjective evaluation using the "Goldilocks" test utterances (two sentences per utterance) synthesised from voices built from either clean or noisy data.

thetic speech utterances from the various systems in a random order. They were then asked to score the naturalness of the utterance using mean opinion score (MOS) on a five point scale, where 5 corresponds to natural and 1 corresponds to unnatural. The listening tests were separateky conducted for each story. In the "Goldilocks" story the systems using different amount of speech data above were evaluated together.

Figure 3 shows the mean opinion scores, with 95% confidence intervals, using the clean speech data for the "Little Girl" utterances. From this result, we can see that the hybrid system is rated as the most natural. This result is consistent with previous findings [8]. Figure 2 shows the mean opinion scores, with 95% confidence intervals, for the "Goldilocks" utterances. Figure 2(a) shows the results for clean speech data, and 2(b) shows the results for noisy speech data. Comparing Figure 2(a) and Figure 3, we notice that subjects no longer rate the hybrid system as the most natural. Further work is needed to discover if this is because the test utterances consisted of two sentences, or whether there is some other reason.

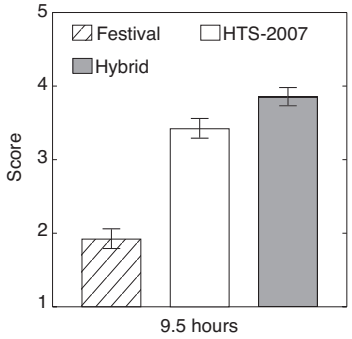


Figure 3: Subjective evaluation using the "Little Girl" test utterances (one sentence per utterance) synthesised from voices built from clean data.

Comparing Figures 2(a) and 2(b), we notice first that the unit-selection method is very poor on noisy data. This is because inconsistency in the recording conditions from session to session translates into inconsistency in the synthetic speech from unit to unit, which makes the resulting synthetic speech "patchy" and very unnatural-sounding. The hybrid system suffers from the the same problem to some extent, since it also concatenates waveforms to generate speech. The speaker-adaptive HTS-2007 system is clearly the most robust of the systems: its performance is least degraded by the use of noisy data. The naturalness of the HTS-2007 system increases as more data become available, although the other systems are unable to improve naturalness by using more data. We believe that there are two principal reasons for the superior robustness of the speaker-adaptive HTS-2007 system. The first is that the average voice model is trained from a large amount of clean speech data. Therefore, the decision trees used for tying of HMM parameters are not affected by the noisy data at all. The second is that the speaker adaptation algorithms used in the system include feature transforms. These feature transform can be viewed as a generalisation of several normalisation techniques such as cepstral mean normalisation, cepstral variance normalisation (CVN), stochastic matching, bias removal and so on. They can normalise the fluctuations of the recording conditions, assuming that these can be approximated by linear or piecewise linear regression.

Our results therefore demonstrate a newly-discovered significant advantage of HMM-based speech synthesis (especially speaker-adaptive HMM-based speech synthesis): "robustness." This ability to generate a synthetic voice from noisy data further expands the potential applications of this technique.

4. Recording Condition-adaptive Training

4.1. Framework

From the positive results above, we propose a new robust training method in which we explicitly aim to normalise the variation of the recording condition. The above results imply the effectiveness of a feature transform for this purpose. Therefore, we propose an algorithm that is analogous to speaker adaptive training (SAT) [22] except the feature transform classes reflect not speakers but recording sessions. In the feature-space SAT algorithm, it is assumed that each state of the HMM has the following Gaussian pdf $b_i(\boldsymbol{o}_{t_r})$, characterised by a mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$:

$$b_i(\boldsymbol{o}_{t_r}) = |\boldsymbol{\zeta}_r| \mathcal{N}(\boldsymbol{\zeta}_r \boldsymbol{o}_{t_r} + \boldsymbol{\epsilon}_r; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (1)$$

where $\boldsymbol{o}_{t_r}$ is the observation vector at time t in recording session r, and $(\boldsymbol{\zeta}_r, \boldsymbol{\epsilon}_r)$ is the set of linear transforms which normalise the feature vector in the recording session r. The set of parameters $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\zeta}_r, \boldsymbol{\epsilon}_r)$ which locally maximises a likelihood function of the training data is used in the following experiment.

4.2. Evaluation

We built "Nitech-HTS 2005" systems [23], which are speaker-dependent versions of the HTS-2007 system, using this method. The speech data used for training the HMMs was the same four hours of speech data used in the noisy condition experiment reported earlier. The systems built from the data with or without the recording condition-adaptive training were then used to synthesise the opening paragraph of "The Emperor's New Clothes" by Hans Christian Anderson. We ask interested readers to listen to the audio examples available at `http://homepages.inf.ed.ac.uk/jyamagis/interspeech08` themselves to hear the differences between these systems.

In a small informal listening test, we confirmed that the system using the recording condition-adaptive training can generate significantly more stable synthetic speech. A more detailed evaluation and analysis, plus integration into the speaker-adaptive approach, as used in the main experiments reported here, is left as future work. The proposed system would involve simultaneous normalisation of the multiple source speakers and different recording sessions when training the Average Voice model. This should improve quality: the multi-speaker training data we use are typically drawn from several corpora recorded under differing conditions in various studios.

5. Conclusions

Whilst the Blizzard Challenge has provided useful comparisons of various speech synthesis methods using clean, phonetically balanced speech data, it has not considered what happens when the speech data are not perfectly clean, recording conditions are not consistent, and/or the phonetic balance of the texts are not ideal. In this paper, we compared the performance of several speech synthesis methods under both clean and noisy conditions. We conclude that speaker-adaptive HMM-based speech synthesis is far more robust than either concatenative, speaker-dependent HMM-based, or hybrid approaches. We have also proposed a new adaptive training method for HMM-based speech synthesis.

6. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project). The work of Z. Ling was supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

This work has made use of the resources provided by the Edinburgh Compute and Data Faclity (ECDF). (`http://www.ecdf.ed.ac.uk/`). The ECDF is partially supported by the eDIKT initiative. (`http://www.edikt.org`.)

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH-99*, Sep. 1999, pp. 2374–2350.

[2] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[3] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis,," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.

[4] M. Aylett and J. Yamagishi, "Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning," in *Proc. LangTech 2008*, Feb. 2008.

[5] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP-96*, May 1996, pp. 373–376.

[6] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments," in *Proc. Interspeech 2006*, Sep. 2006, pp. 2034–2037.

[7] ——, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," in *Proc. ICASSP 2007*, Apr. 2007, pp. 1245–1248.

[8] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Y. J. Chen, and G.-P. Hu, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[9] A. Black, P. Taylor, and R. Caley, *The Festival Speech Synthesis System*, University of Edinburgh, 1999.

[10] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[11] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[12] J. Yamagishi, T. Nose, H. Zen, T. Toda, and K. Tokuda, "Performance evaluation of the speaker-independent HMM-based speech synthesis system HTS-2007 for the Blizzard Challenge 2007," in *Proc. ICASSP 2008*, Apr. 2008.

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, 2007, (accept).

[14] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[15] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[16] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.0.1*, http://hts.sp.nitech.ac.jp/.

[17] Y. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP 2006*, May 2006, pp. 89–92.

[18] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.

[19] H. Kawahara, H. Katayose, A. Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. EUROSPEECH 1999*, Sep. 1999, pp. 2781–2784.

[20] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.

[21] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech 1999*, vol. 2, Sep. 1999, pp. 823–826.

[22] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[23] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.